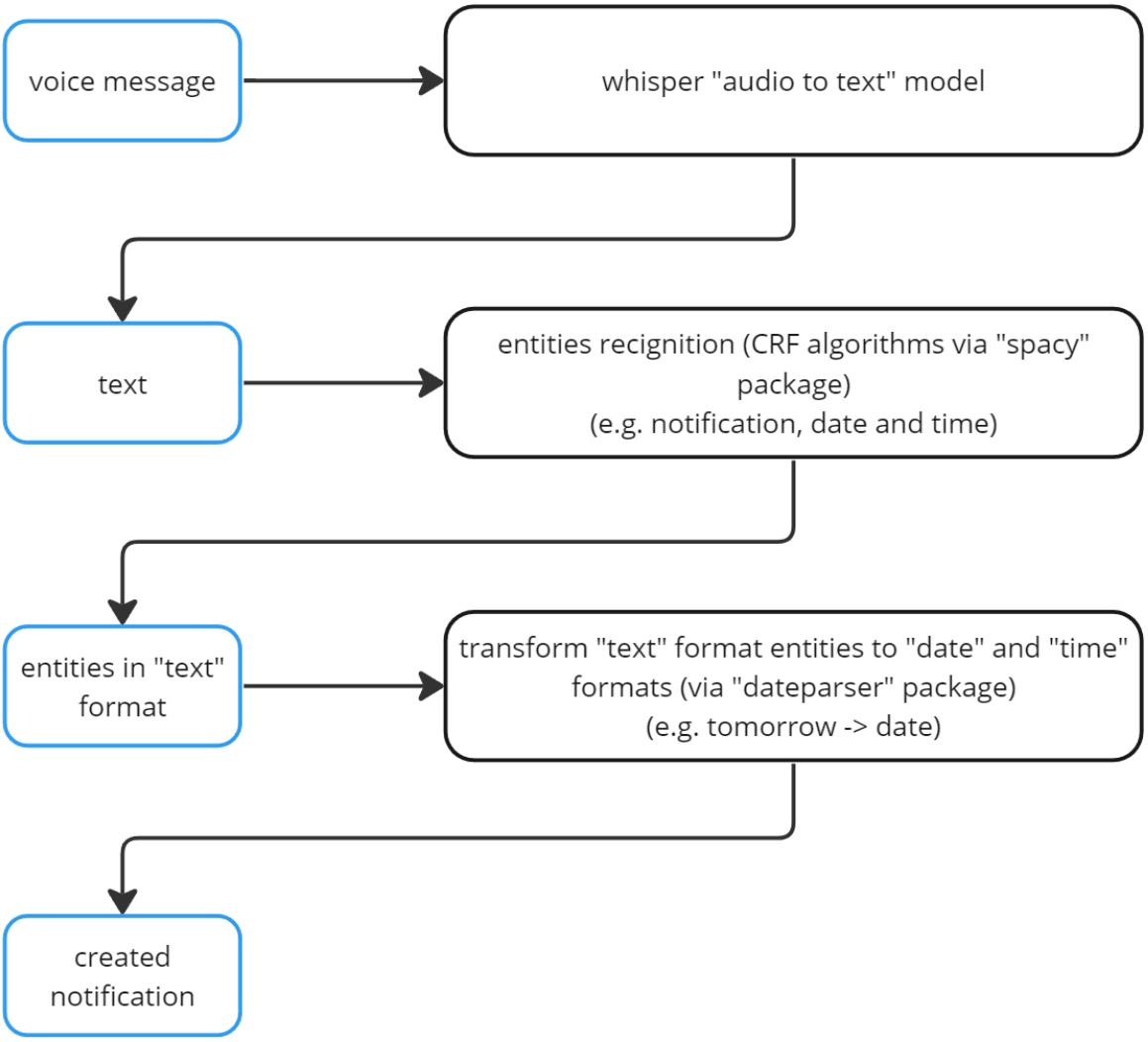
DOCUMENTATION

1) Program logic



2) How spaCy implements NER

spaCy's NER model is based on a machine learning algorithm known as a conditional random field (CRF). The model takes in a sequence of words (tokens) as input and outputs a sequence of labels indicating whether each token is part of a named entity and, if so, which type of entity it belongs to.

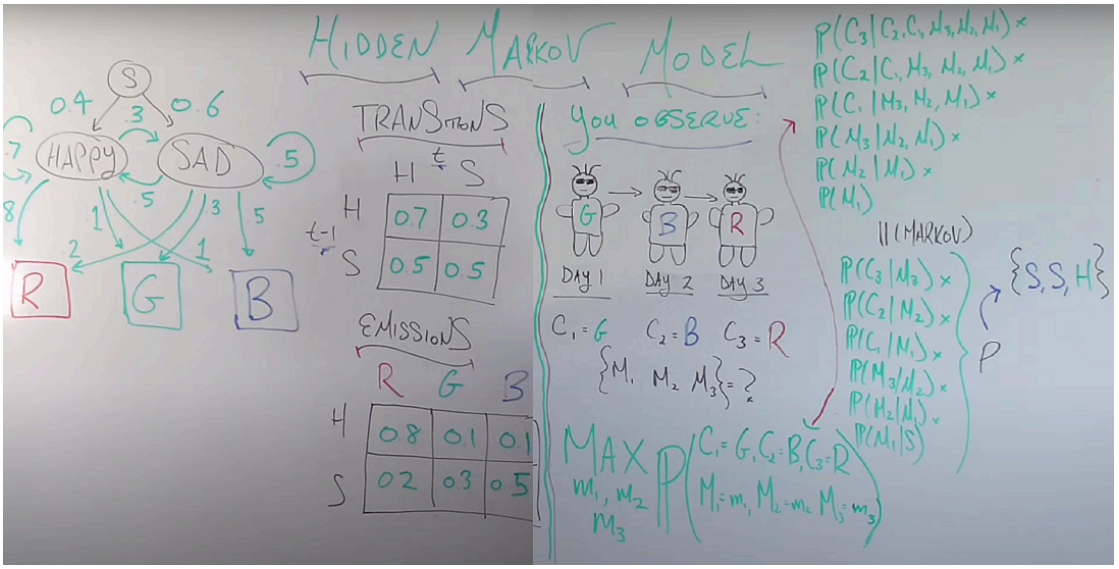
spaCy's NER model is pre-trained on a large corpus of text and can recognize a wide range of named entities out-of-the-box, including people, organizations, locations, and dates. The model also includes additional entity types, such as product names, languages, and nationalities.

3) How CRF works

Actually, CRF is the generalization of the Hidden Markov Model (HMM).

a) HMM intuitive explanation (Example)

https://www.youtube.com/watch?v=fX5bYmnHqqE&list=LL&index=34



- hidden entities: moods {happy, sad}. We should guess them by:
- t-shirt color: {R, G, B}. We can see the sequence of them (e.g. during 3 days)

- transitions matrix we get after learning
- emissions matrix we get after learning

- prediction: maximizing the target function:

$$max_{m1,m2,m3} P(C_1 = R, C_2 = G, C_3 = B, M_1 = m_1, M_2 = m_2, M_3 = m_3)$$

- if we make 2 assumptions: colors depend only on layers; current state depends only on previous (Markov condition) => we can compute the value quickly.

b) **HMM mathematical explanation**

https://en.wikipedia.org/wiki/Hidden_Markov_model

c) **CRF implementation**

https://www.youtube.com/watch?v=rI3DQS0P2fk&list=LL&index=31

Actually, everything is the same except assumptions => more complicated computations

4) Results analysis

a) Metrics (for example for "NTFY" entity:

i) TP: the number of items that marked as "NTFY" in X_test dataset and exist in y_test
ii) FP: the number of items that marked as "NTFY" in X_test dataset and do NOT exist in y_test
iii) FN: the number of items in y_test that marked as "NTFY" and do NOT exist in X_test
iv) TN the number of items in y_test that do NOT marked as "NTFY" and do NOT exist in X_test

precision = TP / (TP + FP)
recall = TP / (TP + FN)

b) Results

i) We have approximately 150 items in train dataset and 20 items in test dataset

ii) Entities metrics:

```
NTFY
precision 0.9482758620689655
recall 0.9016393442622951
TIME
precision 0.967741935483871
recall 0.9523809523809523
DATE
precision 0.9879518072289156
recall 0.9111111111111111
```

c) Progress steps

i) Problems:

NTFY: sometimes takes dates in account, do not treat "please", do not contain Geographical items

DATE: in a week is treated as "a week ago", does not understand "next week", works badly with explicit dates like "5th of…"

TIME: in 30 minutes treats as "30 min ago"

ii) Updated dataset, added new items, enlarged the number of items and the number of epochs (from 20 to 30). Problems are solved.

5) Working examples

a) Remind me today at six thirty to visit my friends
b) I plan to visit France tomorrow at 7am
c) Feed my cat the next week
d) Remind me to send a message on the April 5th, 2024 at 12:25
e) Please, check the bank account tomorrow at 3pm
f) Remind me to walk with my dog in 30 minutes

6) Notes

a) be careful while testing on your own sentences. They must have dots (.) in the end.

7) Auxiliary feature with text summarization

a) Use Case: recording news or lecture parts to have a note about them in a summarized way. Translates 1-5 minutes records to short, 3-5-sentence summarizations.

b) Pre-trained model: BART (large-sized model), fine-tuned on CNN Daily Mail

c) Training dataset: BBC News Summary

d) Model params:

   i) max_seq_length = 400
   ii) min_seq_length = 100 (so, if send too short voice message, it is sends not summarized)
   iii) train_epochs = 10

e) Notebooks:

   UntrainedBART

   TrainedBART

f) Clean model scores:

   ROUGE: 0.18305898037545495
   BLUE: 0.06847058799520754

g) Trained model scores:

ROUGE: 0.46673483193181753
BLUE: 0.44376315777282743

h) Example. Article

Ad sales boost Time Warner profit

Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (£600m) for the three months to December, from $639m year-earlier.

The firm, which is now one of the biggest investors in Google, benefited from sales of high-speed internet connections and higher advert sales. TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn. Its profits were buoyed by one-off gains which offset a profit dip at Warner Bros, and less users for AOL.

Time Warner said on Friday that it now owns 8% of search-engine Google. But its own internet business, AOL, had has mixed fortunes. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding three quarters. However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues. It hopes to increase subscribers by offering the online service free to TimeWarner internet customers and will try to sign up AOL's existing customers for high-speed broadband. TimeWarner also has to restate 2000 and 2003 results following a probe by the US Securities Exchange Commission (SEC), which is close to concluding.

Time Warner's fourth quarter profits were slightly better than analysts' expectations. But its film division saw profits slump 27% to $284m, helped by box-office flops Alexander and Catwoman, a sharp contrast to year-earlier, when the third and final film in the Lord of the Rings trilogy boosted results. For the full-year, TimeWarner posted a profit of $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn. "Our financial performance was strong, meeting or exceeding all of our full-year objectives and greatly enhancing our flexibility," chairman and chief executive Richard Parsons said. For 2005, TimeWarner is projecting operating earnings growth of around 5%, and also expects higher revenue and wider profit margins.

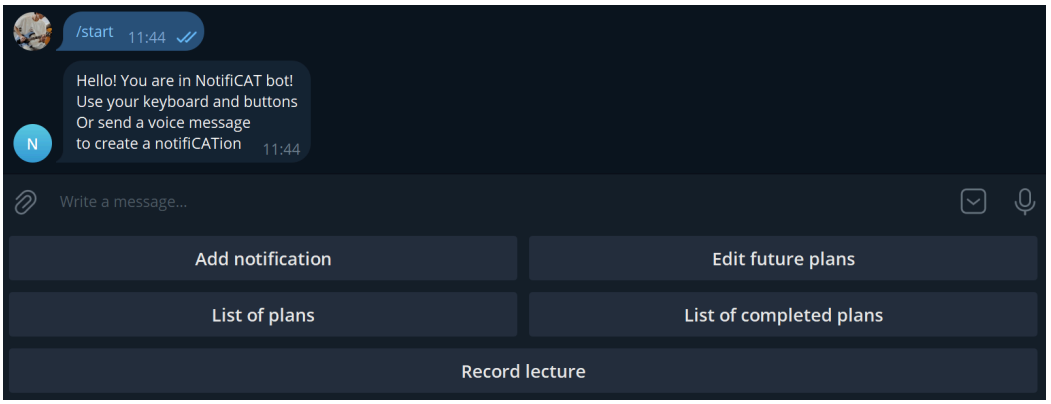
TimeWarner is to restate its accounts as part of efforts to resolve an inquiry into AOL by US market regulators. It has already offered to pay $300m to settle charges, in a deal that is under review by the SEC. The company said it was unable to estimate the amount it needed to set aside for legal reserves, which it previously set at $500m. It intends to adjust the way it accounts for a deal with German music publisher Bertelsmann's purchase of a stake in AOL Europe, which it had reported as advertising revenue. It will now book the sale of its stake in AOL Europe as a loss on the value of that stake.

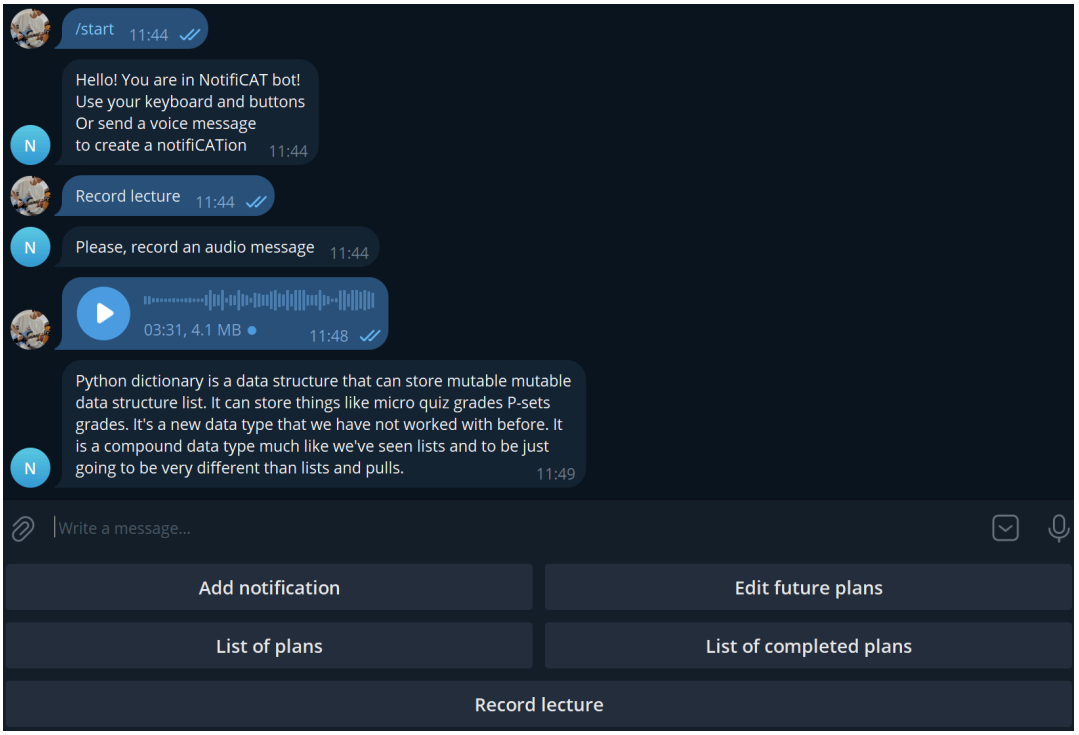
i) Example. Summary

TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn.However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues.Time Warner said on Friday that it now owns 8% of search-engine Google.TimeWarNER's fourth quarter profits were slightly better than analysts' expectations.For the full-year, TimeWarner posted a profit of $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to £42.09bn.The firm, which is now one of the biggest investors in Google, benefited from sales of high-speed internet connections and higher advert sales.However its own internet business, AOL, had has mixed fortunes.

j) Pipeline example in telegram

Link to the lecture




/start 11:44
Hello! You are in NotifiCAT bot!
Use your keyboard and buttons
Or send a voice message
to create a notifiCATion 11:44
Record lecture 11:44
Please, record an audio message 11:44
03:31, 4.1 MB 11:48
Python dictionary is a data structure that can store mutable mutable data structure list. It can store things like micro quiz grades P-sets grades. It's a new data type that we have not worked with before. It is a compound data type much like we've seen lists and to be just going to be very different than lists and pulls. 11:49
Write a message...
Add notification
Edit future plans
List of plans
List of completed plans
Record lecture